A study on the impact of fuzzification in classification and prediction of plants using machine learning models

T. Swathi and S. Sudha*
Department of Electrical and Electronics Engineering, National Institute of Technology, Tiruchirappalli 620 015, India

The objective of the present study is to explore the impact of fuzzification in improving the classification and prediction of plants. The process of transforming numerical values into fuzzified values, thereby allowing uncertainty and imprecision in the model, is known as fuzzification. This study utilises the crop dataset, which has 10 plant types, each with 9 features collected from Kaggle for fuzzification. This fuzzified dataset is trained and tested with various machine learning models such as support vector machine (SVM), Naïve Bayes, random forest, decision tree, XGBoost, K-nearest neighbour (KNN) and LightGBM. These machine learning models are also used for non-fuzzified dataset training and testing. For performance enhancement, the hyperparameters of the machine learning models are fine-tuned using Bayesian Optimization. Performance comparison is based on the evaluation metrics such as accuracy, area under the curve (AUC), precision, recall and F1-score. Results indicate that SVM and KNN benefit significantly due to fuzzification. The SVM accuracy using fuzzified and non-fuzzified data is found to improve from 85.84% to 90.29%, AUC from 98.93% to 99.36%, and F1-score from 86% to 90% respectively. Similarly, for KNN, accuracy is also found to increase from 80.18% to 88.24%, AUC from 97.81% to 99.18%, and F1-score from 80% to 88%. Models like LightGBM and XGBoost are found to consistently maintain high performance across both datasets. The findings support fuzzification's ability to help some models handle complex input better, which results in more accurate categorisation. Therefore, the adaptability of fuzzy models in conjunction with machine learning models demonstrates their use in agricultural applications as well.

Keywords: Accuracy, area under the curve, Bayesian Optimization, classification, cross-validation, fuzzified dataset, prediction.

AGRICULTURE and environmental studies rely heavily on the classification of plant species based on the soil and environmental features. Temperature, moisture, pH, and the nutrient status of the soil influence plant properties. pH determines nutritional availability, while moisture impacts root absorption[1]. Water pollution, nutritional imbalance, leaching, and soil degradation are all results of poor management, particularly excessive fertiliser use. For sustainable farming, this emphasises the necessity of mapping species–soil connections[2]. Due to fluctuating conditions and imperfect measurements, which increase uncertainty, the work is still challenging. In this context, techniques that explicitly describe uncertainty and nonlinear interactions are recommended, as traditional machine learning (ML) algorithms often perform poorly on raw inputs. Advanced ML techniques have demonstrated effectiveness in addressing these challenges, making significant contributions to various domains, including healthcare, finance, and environmental studies.

In healthcare, a diabetes prediction model using ML techniques such as logistic regression (LR), support vector machine (SVM), Naïve Bayes, and random forest (RF) is developed[3]. Ensemble methods, including extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), CatBoost, AdaBoost, and bagging, are employed to enhance prediction accuracy. Among these techniques, high scores are obtained by CatBoost with an accuracy of 95.4%, and an area under the curve-receiver operating characteristic (AUC-ROC) score of 99%, outperforming other ensemble methods. LR, K-nearest neighbour (KNN), and decision tree (DT) were used to classify and predict student depression[4]. A dataset of 787 college students was analysed, and among the models, LR demonstrated the highest accuracy of 77%, along with 70% recall and a 72% F1-score. Similarly, many studies on classification and prediction using ML in other areas are discussed.

In cybersecurity, a ML-based hyperparameter Optimization method is used to categorise network intrusions[5]. The features are trained using ML models, including XGBoost, LightGBM, CatBoost, RF and DT, with hyperparameter tuning to enhance performance. Among these, the LightGBM demonstrated superior performance, achieving a classification accuracy of 99.77%.

In finance, stock price prediction is studied by combining long short-term memory with ML models like RF, SVM and LightGBM[6]. A novel dataset filtering method enhances decision-making for buying and

*For correspondence. (e-mail: sudha@nitt.edu)